Wrangle and Analyze Data Project

Wrangle report

By: Afnan Abdullah K Alshehri

Table of contents

Introdiction:

In this project I will take three steps to wrangle the dataset from the tweet archive of Twitter user @dog_rates, also known as WeRateDogs. WeRateDogs is a Twitter account that rates people's dogs with humorous comments about the dog.

The three steps are:

1.Gathering Data: In this step, I will gather all three pieces of data that we must use.

2. Assessing Data: After gathering all three pieces of data, assess them visually and programmatically for quality and tidiness issues.

3.Cleaning Data:Clean all of the issues I documented while assessing.

Gathering Data:

In this step, I have gathered datasets from different sources:
1.Directly download the WeRateDogs Twitter archive data (twitter_archive_enhanced.csv)
2.Use the Requests library to download the tweet image prediction (image_predictions.tsv)
3.Use the Tweepy library to query additional data via the Twitter API (tweet_json.txt)

Assessing Data:

In this step, I detected and documented many quality issues and tidiness issues. We must use both visual assessment and programmatic assessment to assess the data.

Quality issues:
1.These columns (in_reply_to_status_id, in_reply_to_user_id, retweeted_status_id,retweeted_status_user_id, retweeted_status_timestamp) are not needed and contain many null values.

2.timestamp should be converted to data type.

3.The Anchor element in source.

4.We have 55 dogs whose name is a.

5.source should be converted to category data type.

6.We have 59 nulls in expanded_urls.

7.The denominator greater than 10 and less than 10.

8.img_num should be converted to category data type.

9.Underscores in p1 and p2 and p3.

10.Inconsistent capitalization in dog breeds in p1,p2 columns.

11.Some column names are not clear.

12.Drop rating_numerator and rating_denominator.

13.Drop doggo and floofer and pupper,puppo columns.

Tidiness issues:
1. In twitter archive dataset we should have one column instead two columns rating_numerator and rating_denominator by divided rating_numerator and rating_denominator

2. In twitter archive dataset we should have one column for dog type instead four columns
3. merge all dataframe in one frame

3.Cleaning Data:

In this step, I clean all of the issues I documented while assessing.

During cleaning, I use the define-code-test framework.